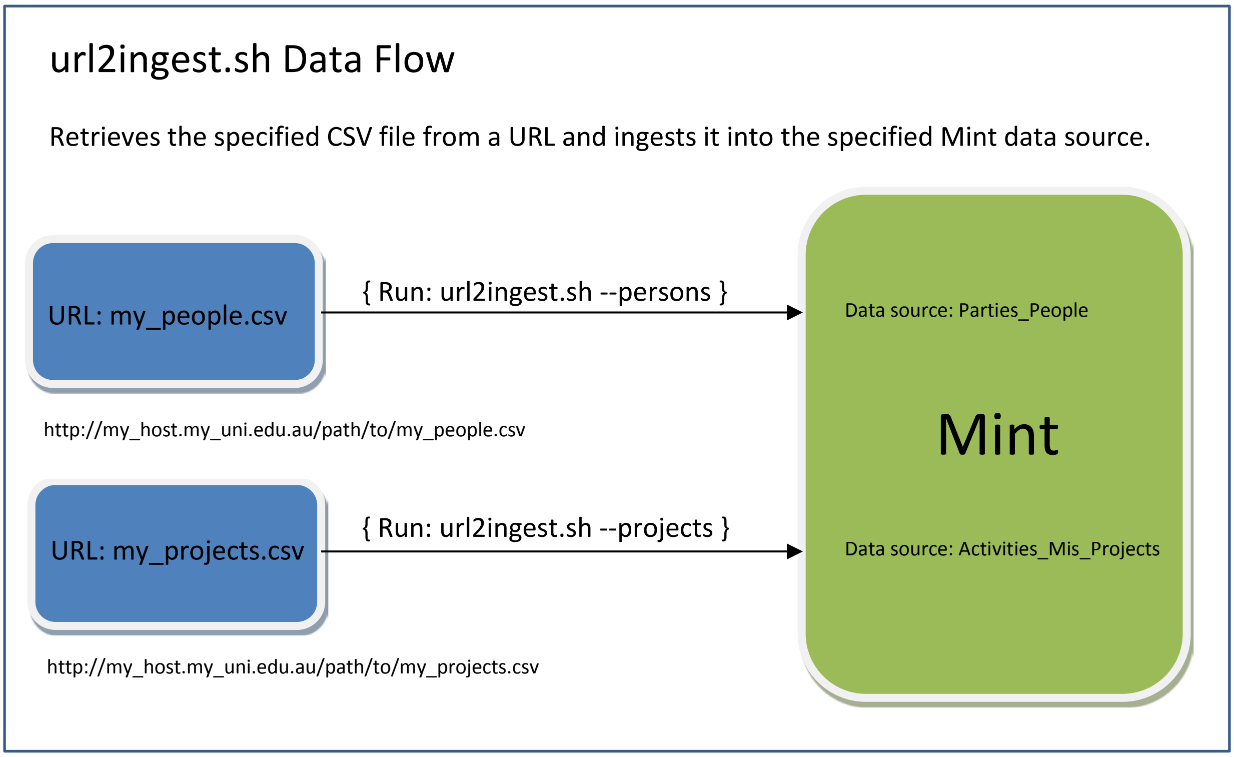
FlindersRedbox-url2ingest: Development documentation

See the INSTALL document for information regarding application environment, installation and configuration.

Although the script is written in bash, it is expected that it will operate with little or no modification under sh and ksh shells.

Data Flow

Unless incremental-load is specified on the command line, there is no data transformation performed by the script. If incremental-load is specified (not shown in the diagram below) the data loaded into Mint is filtered to leave only new or updated records. The script data flow diagram is shown below.



url2ingest.sh Data Flow

Retrieves the specified CSV file from a URL and ingests it into the specified Mint data source.

URL: my_people.csv
{ Run: url2ingest.sh --persons }
http://my_host.my_uni.edu.au/path/to/my_people.csv
URL: my_projects.csv
{ Run: url2ingest.sh --projects }
http://my_host.my_uni.edu.au/path/to/my_projects.csv
Data source: Parties_People
Mint
Data source: Activities_Mis_Projects

High level algorithm

The following high level algorithm is used in the url2ingest.sh script.

```
Read command line arguments and initialise URL, data source, etc variables.
Get CSV file from URL and store in download directory.
Apply an inclusive-filter. [1]

IF full load argument specified THEN
  Load CSV file into Mint data source (if any data records). [2]

ELSEIF incremental load argument specified THEN
  Find the most recent filtered-CSV file.
  Extract new and updated CSV records (compared with most recent filtered-CSV file).

  IF there are any new or updated records THEN
    Load CSV file into Mint data source (if any data records). [2]
  ENDIF

ENDIF

IF the CSV-load was successful THEN
  Backup CSV files.
ENDIF

IF the CSV-load was attempted (successful or not) THEN
  Backup Mint harvest.out file into log directory.
ENDIF
Clean up download directory, working directory, etc.
```

Notes

[1] The inclusive filter can be adjusted by the list of regular expressions in etc/include_filter_people.conf and etc/include_filter_project.conf for persons and projects respectively.

[2] A symbolic link from the CSV file in the Mint tree must be configured in advance to point to the appropriate FINAL_FPATH for the data source. After configuring url2ingest.sh, it can be run with the --dump switch to show FINAL_FPATH for both persons and projects.